(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: DUMAS MILNE EDWARDS, Jean-Baptiste [FR/FR]; 8, rue Grégoire de Tours, F-75006 Paris (FR). BOUGUELERET, Lydie [FR/FR]; 108, avenue Victor Hugo, F-92170 Vanves (FR). JOBERT, Séverin [FR/FR]; 7, impasse Tourneux, F-75010 Paris (FR).

(54) Title: FULL-LENGTH HUMAN cDNAs ENCODING POTENTIALLY SECRETED PROTEINS

(57) Abstract: The invention concerns GENSET polynucleotides and polypeptides. Such GENSET products may be used as reagents in forensic analyses, as chromosome markers, as tissue/cell/organelle-specific markers, in the production of expression vectors. In addition, they may be used in screening and diagnosis assays for abnormal GENSET expression and/or biological activity and for screening compounds that may be used in the treatment of GENSET-related disorders.

informatics software may mischaracterize the genomic sequences obtained, *i.e.*, labeling non-coding DNA as coding DNA and vice versa.

An alternative approach takes a more direct route to identifying and characterizing human genes. In this approach, complementary DNAs (cDNAs) are synthesized from isolated messenger

5   RNAs (mRNAs) which encode human proteins. Using this approach, sequencing is only performed on DNA which is derived from protein coding fragments of the genome. In the past, these cDNAs, ofter short EST sequences were obtained from oligo-dT primed cDNA libraries. Accordingly, they mainly corresponded to the 3' untranslated region of the mRNA. In part, the prevalence of EST sequences derived from the 3' end of the mRNA is a result of the fact that typical techniques for

10 obtaining cDNAs, are not well suited for isolating cDNA sequences derived from the 5' ends of mRNAs (Adams *et al., Nature* 377:3-174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996). In addition, in those reported instances where longer cDNA sequences have been obtained, the reported sequences typically correspond to coding sequences and do not include the full 5' untranslated region (5'UTR) of the mRNA from which the cDNA is derived. Indeed, 5'UTRs have

15 been shown to affect either the stability or translation of mRNAs. Thus, regulation of gene expression may be achieved through the use of alternative 5'UTRs as shown, for instance, for the translation of the tissue inhibitor of metalloprotease mRNA in mitogenically activated cells (Waterhouse *et al., J Biol Chem.* 265:5585-9. 1990). Furthermore, modification of 5'UTR through mutation, insertion or translocation events may even be implied in pathogenesis. For instance, the

20 fragile X syndrome, the most common cause of inherited mental retardation, is partly due to an insertion of multiple CGG trinucleotides in the 5'UTR of the fragile X mRNA resulting in the inhibition of protein synthesis via ribosome stalling (Feng *et al., Science* 268:731-4, 1995). An aberrant mutation in regions of the 5'UTR known to inhibit translation of the proto-oncogene *c-myc* was shown to result in upregulation of c-myc protein levels in cells derived from patients with

25 multiple myelomas (Willis *et al., Curr Top Microbiol Immunol* 224:269-76, 1997). In addition, the use of oligo-dT primed cDNA libraries does not allow the isolation of complete 5'UTRs since such incomplete sequences obtained by this process may not include the first exon of the mRNA, particularly in situations where the first exon is short. Furthermore, they may not include some exons, often short ones, which are located upstream of splicing sites. Thus, there is a need to obtain

30 sequences derived from the 5' ends of mRNAs.

Moreover, despite the great amount of EST data that large-scale sequencing projects have yielded (Adams *et al., Nature* 377:174, 1996, Hillier *et al., Genome Res.* 6:807-828, 1996), information concerning the biological function of the mRNAs corresponding to such obtained cDNAs has revealed to be limited. Indeed, whereas the knowledge of the complete coding

35 sequence is absolutely necessary to investigate the biological function of mRNAs, ESTs yield only partial coding sequences. So far, large-scale full-length cDNA cloning has been achieved only with limited success because of the poor efficiency of methods for constructing full-length cDNA

2

libraries. Indeed, such methods require either a large amount of mRNA (Ederly *et al.*, 1995), thus resulting in non representative full-length libraries when small amounts of tissue are available or require PCR amplification (Maruyama *et al.*, 1994; CLONTECHniques, 1996) to obtain a reasonable number of clones, thus yielding strongly biased cDNA libraries where rare and long

5   cDNAs are lost. Thus, there is a need to obtain full-length cDNAs, *i.e.* cDNAs containing the full coding sequence of their corresponding mRNAs. The present application presents a number of cDNAs, called GENSET polynucleotides, isolated from full-length cDNA librairies obtained from the methods described in PCT publication WO 00/37491.

      While many sequences derived from human chromosomes have practical applications,

10   approaches based on the identification and characterization of those chromosomal sequences which encode a protein product are particularly relevant to diagnostic and therapeutic uses. Of the 50,000-100,000 protein coding genes, those genes encoding proteins which are secreted from the cell in which they are synthesized, as well as the secreted proteins themselves, are particularly valuable as potential therapeutic agents. Such proteins are often involved in cell to cell communication and

15   may be responsible for producing a clinically relevant response in their target cells. In fact, several secretory proteins, including tissue plasminogen activator, G-CSF, GM-CSF, erythropoietin, human growth hormone, insulin, interferon-α, interferon-β, interferon-γ, and interleukin-2, are currently in clinical use. These proteins are used to treat a wide range of conditions, including acute myocardial infarction, acute ischemic stroke, anemia, diabetes, growth hormone deficiency, hepatitis, kidney

20   carcinoma, chemotherapy induced neutropenia and multiple sclerosis. For these reasons, cDNAs encoding secreted proteins or fragments thereof represent a particularly valuable source of therapeutic agents. Thus, there is a need for the identification and characterization of secreted proteins and the nucleic acids encoding them.

      In addition to being therapeutically useful themselves, secretory proteins include short

25   peptides, called signal peptides, at their amino termini which direct their secretion. These signal peptides are encoded by the signal sequences located at the 5' ends of the coding sequences of genes encoding secreted proteins. Because these signal peptides will direct the extracellular secretion of any protein to which they are operably linked, the signal sequences may be exploited to direct the efficient secretion of any protein by operably linking the signal sequences to a gene encoding the

30   protein for which secretion is desired. In addition, fragments of the signal peptides called membrane-translocating sequences, may also be used to direct the intracellular import of a peptide or protein of interest. This may prove beneficial in gene therapy strategies in which it is desired to deliver a particular gene product to cells other than the cells in which it is produced. Signal sequences encoding signal peptides also find application in simplifying protein purification

35   techniques. In such applications, the extracellular secretion of the desired protein greatly facilitates purification by reducing the number of undesired proteins from which the desired protein must be

selected. Thus, there exists a need to identify and characterize the 5' fragments of the genes for secretory proteins which encode signal peptides.

Sequences coding for human proteins may also find application as therapeutics or diagnostics. In particular, such sequences may be used to determine whether an individual is likely
5   to express a detectable phenotype, such as a disease, as a consequence of a mutation in the coding sequence for a protein. In instances where the individual is at risk of suffering from a disease or other undesirable phenotype as a result of a mutation in such a coding sequence, the undesirable phenotype may be corrected by introducing a normal coding sequence using gene therapy. Alternatively, if the undesirable phenotype results from overexpression of the protein encoded by
10   the coding sequence, expression of the protein may be reduced using antisense or triple helix based strategies.

The GENSET human polypeptides encoded by the coding sequences may also be used as therapeutics by administering them directly to an individual having a condition, such as a disease, resulting from a mutation in the sequence encoding the polypeptide. In such an instance, the
15   condition can be cured or ameliorated by administering the polypeptide to the individual.

In addition, the human polypeptides or fragments thereof may be used to generate antibodies useful in determining the tissue type or species of origin of a biological sample. The antibodies may also be used to determine the subcellular localization of the human polypeptides or the cellular localization of polypeptides which have been fused to the human polypeptides. In
20   addition, the antibodies may also be used in immunoaffinity chromatography techniques to isolate, purify, or enrich the human polypeptide or a target polypeptide which has been fused to the human polypeptide.

Public information on the number of human genes for which the promoters and upstream regulatory regions have been identified and characterized is quite limited. In part, this may be due
25   to the difficulty of isolating such regulatory sequences. Upstream regulatory sequences such as transcription factor binding sites are typically too short to be utilized as probes for isolating promoters from human genomic libraries. Recently, some approaches have been developed to isolate human promoters. One of them consists of making a CpG island library (Cross *et al., Nature Genetics* 6: 236-244, 1994). The second consists of isolating human genomic DNA sequences
30   containing SpeI binding sites by the use of SpeI binding protein. (Mortlock *et al., Genome Res.* 6:327-335, 1996). Both of these approaches have their limits due to a lack of specificity and of comprehensiveness. Thus, there exists a need to identify and systematically characterize the 5' fragments of the genes.

cDNAs including the 5' ends of their corresponding mRNA may be used to efficiently
35   identify and isolate 5'UTRs and upstream regulatory regions which control the location, developmental stage, rate, and quantity of protein synthesis, as well as the stability of the mRNA (Theil *et al., BioFactors* 4:87-93, (1993). Once identified and characterized, these regulatory

4

regions may be utilized in gene therapy or protein purification schemes to obtain the desired amount and locations of protein synthesis or to inhibit, reduce, or prevent the synthesis of undesirable gene products.

In addition, cDNAs containing the 5' ends of protein genes may include sequences useful as probes for chromosome mapping and the identification of individuals. Thus, there is a need to identify and characterize the sequences upstream of the 5' coding sequences of genes encoding proteins.

Summary of the invention

The present invention provides compositions containing a purified or isolated polynucleotide comprising, consisting of, or consisting essentially of a nucleotide sequence selected from the group consisting of: (a) the sequences of SEQ ID Nos: 1-241; (b) the sequences of clone inserts of the deposited clone pool; (c) the full coding sequences of SEQ ID Nos: 1-241; (d) the full coding sequences of the clone inserts of the deposited clone pool; (e) the sequences encoding one of the polypeptides of SEQ ID Nos: 242-482; (f) the sequences encoding one of the polypeptides encoded by the clone inserts of the deposited clone pool; (g) the genomic sequences coding for GENSET polypeptides; (h) the 5' transcriptional regulatory regions of GENSET genes; (i) the 3' transcriptional regulatory regions of GENSET genes; (j) the polynucleotides comprising the nucleotide sequence of any combination of (g)-(i); (k) the variant polynucleotides of any of the polynucleotides of (a)-(j); (l) the polynucleotides comprising a nucleotide sequence of (a)-(k), wherein the polynucleotide is single stranded, double stranded, or a portion is single stranded and a portion is double stranded; (m) the polynucleotides comprising a nucleotide sequence complementary to any of the single stranded polynucleotides of (l). The invention further provides for fragments of the nucleic acid molecules of (a)-(m) described above.

The present invention also provides biologically active forms, variants, fragments and derivatives of the present proteins, where "biologically active" indicates that the form, variant, fragment, or derivative, has any detectable activity in any in vitro assay known in the art or described herein, or has any detectable function in vivo. In preferred embodiments, a determination of whether a particular polypeptide is biologically active will be made based on any of the specific assays or functional characteristics provided below for each of the proteins of this invention.

Therefore, one embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool, sequences complementary thereto, allelic variants thereof, and degenerate variants thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid comprising at least 8 consecutive nucleotides of a sequence selected from the

group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the
deposited clone pool, sequences complementary thereto, allelic variants thereof, and degenerate
variants thereof. In one aspect of this embodiment, the nucleic acid comprises at least 10, 12, 15,
18, 20, 25, 28, 30, 35, 40, 50, 75, 100, 150, 200, 300, 400, 500, 800, 1000, 1500, or 2000

5   consecutive nucleotides of said selected sequence, sequences complentary thereto, allelic variants
thereof, and degenerate variants thereof. The nucleic acid may be a recombinant nucleic acid.

    Another embodiment of the present invention is a composition comprising a vertebrate
purified or isolated nucleic acid of at least 15,18, 20, 23, 25, 28, 30, 35, 40, 50, 75, 100, 200, 300,
500, 1000 or 2000 nucleotides in length which hybridizes under stringent conditions to any

10   polynucleotide of the invention, preferably a sequence selected from the group consisting of
sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool,
sequences complementary thereto. In one aspect of this embodiment, the nucleic acid is
recombinant.

    Another embodiment of the present invention is a composition containing a purified or

15   isolated nucleic acid comprising the full coding sequences of a sequence selected from the group
consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited
clone pool, or an allelic variant thereof. In one aspect of this embodiment, the nucleic acid is
recombinant.

    A further embodiment of the present invention is a composition containing a purified or

20   isolated nucleic acid comprising a contiguous span of a sequence selected from the group consisting
of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted
proteins in the deposited clone pool, or an allelic variant thereof, wherein said contiguous span
encodes a mature protein. In one aspect of this embodiment, the nucleic acid is recombinant. In
another aspect of this embodiment, the nucleic acid is an expression vector wherein said contiguous

25   span which encodes a mature protein is operably linked to a promoter.

    Yet another embodiment of the present invention is a composition containing a purified or
isolated nucleic acid comprising a contiguous span of a sequence selected from the group consisting
of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted
proteins in the deposited clone pool, or an allelic variant thereof, wherein said contiguous span

30   encodes a signal peptide. In one aspect of this embodiment, the nucleic acid is recombinant. In
another aspect of this embodiment, the nucleic acid is an fusion vector wherein said contiguous
span which encodes a signal peptide is operably linked to a second nucleic acid encoding an
heterologous polypeptide.

    Another embodiment of the present invention is a composition containing a purified or

35   isolated nucleic acid encoding a polypeptide comprising a sequence selected from the group
consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited

clone pool, or allelic variant thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid encoding a polypeptide comprising the sequence of a mature protein included

5	in a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or allelic variant thereof. In one aspect of this embodiment, the nucleic acid is recombinant.

Another embodiment of the present invention is a composition containing a purified or isolated nucleic acid encoding a polypeptide comprising the sequence of a signal peptide included

10	in a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts encoding secreted proteins in the deposited clone pool, or allelic variant thereof. In another aspect it is present in a vector of the invention.

Further embodiments of the invention include compositions containing purified or isolated polynucleotides that comprise, a nucleotide sequence at least 70% identical, more preferably at least

15	75% identical, and still more preferably at least 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identical to any of the polynucleotides of the present invention. Methods of determining identity include those well known in the art and described herein. Such analyses can be performed using a full length polynucleotide sequence or using a subsequence of any length. For example, any two sequences can be compared over a region, in either protein or in both proteins, of any 10, 25, 50,

20	100, 250, 500, 1000, 2000 or more contiguous nucleotides. In addition, any two sequences can be identified as homologous even when they share sequence homology over a limited region of either polynucleotide, for example over a region of at least about 10, 25, 50, 100, 250, 500, 1000, or more contiguous nucleotides.

The invention further provides compositions containing a purified or isolated polypeptide

25	comprising, consisting of, or consisting essentially of an amino acid sequence selected from the group consisting of: (a) the polypeptides of SEQ ID Nos: 242-482; (b) the polypeptides encoded by the clone inserts of the deposited clone pool; (c) the epitope-bearing fragments of the polypeptides of SEQ ID Nos: 242-482; (d) the epitope-bearing fragments of the polypeptides encoded by the clone inserts contained in the deposited clone pool; (e) the domains of the polypeptides of SEQ ID

30	Nos: 242-482; (f) the domains of the polypeptides encoded by the clone inserts contained in the deposited clone pool; and (g) the allelic variant polypeptides of any of the polypeptides of (a)-(f). The invention further provides for fragments of the polypeptides of (a)-(g) above, such as those having biological activity or comprising biologically functional domain(s).

Yet another embodiment of the present invention is a composition containing a purified or

35	isolated protein comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allelic variant thereof.

7

Another embodiment of the present invention is a composition containing a purified or isolated polypeptide comprising at least 5, 6 or 8 consecutive amino acids of a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allellic var:ant thereof. In one aspect of this

5   embodiment, the purified or isolated polypeptide comprises at least 10, 12, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 150, 200, 250, 300, 350, 400, 450 or 500 consecutive amino acids of said selected sequence or allelic variant thereof.

Another embodiment of the present invention is a composition containing an isolated or purified polypeptide comprising a signal peptide of a sequence selected from the group consisting

10   of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, or allellic variant thereof.

Yet another embodiment of the present invention is a composition containing an isolated or purified polypeptide comprising a mature protein of a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by

15   clone inserts of the deposited clone pool, or allellic variant thereof.

A further embodiment of the present invention are compositions containing polypeptide having an amino acid sequence with at least 70% similarity, and more preferably at least 75%, 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% similarity to a polypeptide of the present invention, as well as polypeptides having an amino acid sequence at least 70% identical, more preferably at least

20   75% identical, and still more preferably 80%, 85%, 90%, 95%, 96%, 97%, 98%, or 99% identical to a polypeptide of the present invention. Such analyses can be performed using a full length polypeptide sequence or using a subsequence of any length. For example, any two sequences can be compared over a region, in either protein or in both proteins, of any 10, 25, 50, 100, 250, 500, 1000, 2000 or more contiguous amino acids. In addition, any two sequences can be identified as

25   homologous even when they share sequence homology over a limited region of either protein, for example over a region of at least about 10, 25, 50, 100, 250, 500, 1000, or more contiguous amino acids. Further included in the invention are compositions comprising a purified or isolated nucleic acid molecule encoding such polypeptides. Methods for determining identity include those well known in the art and described herein.

30   The present invention also relates to compositions comprising recombinant vectors, which include the purified or isolated polynucleotides of the present invertion, and to host cells recombinant for the polynucleotides of the present invention, as well as to methods of making such vectors and host cells. The present invention further relates to the use of these recombinant vectors and recombinant host cells in the production of GENSET polypeptides.

35   Consequently, another embodiment of the invention is a vector comprising any polynucleotide of the invention. In a preferred embodiment, the vector is an expression vector comprising a nucleic acid sequence encoding a polypeptide selected from the group consisting of

sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by the clone inserts of the deposited clone pool, or allelic variant thereof, wherein said nucleic acid sequence is operably linked to a promoter. In another preferred embodiment, the vector is a secretion vector comprising a nucleic acid sequence encoding a signal peptide selected from the group consisting of signal

5   peptides of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of secreted polypeptides encoded by the clone inserts of the deposited clone pool, or allelic variant thereof, wherein said nucleic acid sequence is operably linked to an heterologous protein such that said signal peptide will direct the secretion of said heterolgous protein.

A further embodiment of the present invention is a method of making a protein comprising

10 a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, comprising the steps of

a) obtaining a cDNA comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool;

15   b) inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter; and

c) introducing said expression vector into a host cell whereby the host cell produces the protein encoded by said cDNA.

In one aspect of this embodiment, the method further comprises the step of isolating said

20 protein.

Another embodiment of the present invention is a protein obtainable by the method described in the preceding paragraph.

Another embodiment of the present invention is a method of making a protein comprising the amino acid sequence of the mature protein contained in a sequence selected from the group

25 consisting of sequences of SEQ ID NOs: 242-272 and 274-384 and sequences of polypeptides encoded by clone inserts of the deposited clone pool, comprising the steps of

a) obtaining a cDNA comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said cDNA encodes a mature protein;

30   b) inserting said cDNA in an expression vector such that said cDNA is operably linked to a promoter; and

c) introducing said expression vector into a host cell whereby the host cell produces the mature protein encoded by said cDNA.

In one aspect of this embodiment, the method further comprises the step of isolating said

35 protein.

Another embodiment of the present invention is a mature protein obtainable by the method described in the preceding paragraph.

9

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool or a sequence complementary thereto described herein.

5   Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising the full coding sequences of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-241 and sequences of clone inserts of the deposited clone pool.

Another embodiment of the present invention is a composition containing a host cell

10   containing the purified or isolated nucleic acids comprising a contiguous span of a sequence selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said contiguous span codes for a mature protein.

Another embodiment of the present invention is a composition containing a host cell containing the purified or isolated nucleic acids comprising a contiguous span of a sequence

15   selected from the group consisting of sequences of SEQ ID NOs: 1-31 and 33-143 and sequences of clone inserts of the deposited clone pool, wherein said contiguous span codes for a signal peptide.

The invention further relates to other methods of making the polypeptides of the present invention.

The present invention further relates to transgenic plants or animals, wherein said transgenic

20   plant or animal is transgenic for a polynucleotide of the present invention and expresses a polypeptide of the present invention.

The invention further relates to compositions comprising antibodies that specifically bind to the GENSET polypeptides of the present invention and fragments thereof as well as to methods for producing such antibodies and fragments thereof.

25   Therefore, another embodiment of the present invention is a composition containing a purified or isolated antibody capable of specifically binding to a protein comprising a sequence selected from the group consisting of sequences of SEQ ID NOs: 242-482 and sequences of polypeptides encoded by clone inserts of the deposited clone pool. In one aspect of this embodiment, the antibody is capable of binding to a polypeptide comprising at least 6 consecutive

30   amino acids, at least 8 consecutive amino acids, or at least 10 consecutive amino acids of said selected sequence.

The invention also provides kits and methods of detecting GENSET gene expression and/or biological activity in a biological sample. One such method involves assaying for the expression of a GENSET polynucleotide in a biological sample using polymerase chain reaction (PCR) to amplify

35   and detect GENSET polynucleotides or Southern and Northern blot hybridization to detect GENSET genomic DNA, cDNA or mRNA. Alternatively, a method of detecting GENSET gene

expression in a test sample can be accomplished using a compound which binds to a GENSET polypeptide of the present invention or a portion of a GENSET polypeptide.

The present invention also relates to diagnostic methods of identifying individuals or non-human animals having elevated or reduced levels of GENSET products, which individuals are

5   likely to benefit from therapies to suppress or enhance GENSET gene expression, respectively and to methods of identifying individuals or non-human animals at increased risk for developing, or present state of having, certain diseases/disorders associated with GENSET gene abnormal expression or biological activity.

The present invention also relates to kits and methods of screening compounds for their

10   ability to modulate (e.g. increase or inhibit) the activity or expression of GENSET genes including compounds that interact with GENSET gene regulatory sequences and compounds that interact directly or indirectly with GENSET polypeptides. Uses of such compounds are also under the scope of the present invention.

The present invention also relates to pharmaceutical or physiologically acceptable

15   compositions comprising, an active agent, the polypeptides, polynucleotides or antibodies of the present invention.

The present invention also relates to computer systems containing cDNA codes and polypeptides codes of sequences of the invention and to computer-related methods of comparing sequences, identifying homology or features using GENSET sequences of the invention.

20   In another aspect, the present invention provides an isolated polynucleotide, said polynucleotide comprising a nucleic acid sequence encoding i) a polypeptide comprising an amino acid sequence having at least about 80% identity to any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool; or a biologically active fragment of said polypeptide.

25   In one embodiment, the polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of the polypeptides encoded by the clone inserts of the deposited clone pool. In another embodiment, the polypeptide comprises a signal peptide. In another embodiment, the polypeptide is a mature protein. In another embodiment, the nucleic acid sequence has at least about 80% identity over at least about 100 contiguous nucleotides to any one

30   of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide hybridizes under stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the nucleic acid sequence comprises any one of the sequences shown as SEQ ID

35   NOs:1-241 or any one the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide is operably linked to a promoter.

In another aspect, the present invention provides an expression vector comprising the polynucleotide operably linked to a promoter. In another aspect, the present invention provides a host cell recombinant for the polynucleotide. In another aspect, the present invention provides a non-human transgenic animal comprising the host cell.

5   In another aspect, the present invention provides a method of making a GENSET polypeptide, the method comprising a) providing a population of host cells comprising a herein-described polynucleotide and b) culturing the population of host cells under conditions conducive to the production of the polypeptide within said host cells.

In one embodiment, the method further comprises purifying the polypeptide from the 10 population of host cells.

In another aspect, the present invention provides a method of making a GENSET polypeptide, the method comprising a) providing a population of cells comprising a herein-described polynucleotide; b) culturing the population of cells under conditions conducive to the production of the polypeptide within the cells; and c) purifying the polypeptide from the population 15 of cells.

In another aspect, the present invention provides an isolated polynucleotide, the polynucleotide comprising a nucleic acid sequence having at least about 80% identity over at least about 100 contiguous nucleotides to any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

20   In one embodiment, the polynucleotide hybridizes under stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool. In another embodiment, the polynucleotide comprises any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

25   In another aspect, the present invention provides a biologically active polypeptide encoded by any of the herein-described polynucleotides.

In another aspect, the present invention provides an isolated polypeptide or biologically active fragment thereof, the polypeptide comprising an amino acid sequence having at least about 80% sequence identity to any one of the sequences shown as SEQ ID NOs:242-482 or any one of 30 the sequences of polypeptides encoded by the clone inserts of the deposited clone pool.

In one embodiment, the polypeptide is selectively recognized by an antibody raised against an antigenic polypeptide, or an antigenic fragment thereof, said antigenic polypeptide comprising any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool. In another embodiment, the 35 polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool. In another

embodiment, the polypeptide comprises a signal peptide. In another embodiment, the polypeptide is a mature protein.

In another aspect, the present invention provides an antibody that specifically binds to any of ther herein-described polypeptides.

5   In another aspect, the present invention provides a method of determining whether a GENSET gene is expressed within a mammal, the method comprising the steps of: a) providing a biological sample from said mammal; b) contacting said biologica. sample with either of: i) a polynucleotide that hybridizes under stringent conditions to the polynucleotide of claim 1; or ii) a polypeptide that specifically binds to the polypeptide of claim 19; and c) detecting the presence or

10   absence of hybridization between the polynucleotide and an RNA species within the sample, or the presence or absence of binding of the polypeptide to a protein within the sample; wherein a detection of the hybridization or of the binding indicates that the GENSET gene is expressed within the mammal.

In one embodiment, the polynucleotide is a primer, and the hybridization is detected by

15   detecting the presence of an amplification product comprising the sequence of the primer. In another embodiment, the polypeptide is an antibody.

In another aspect, the present invention provides a method of determining whether a mammal has an elevated or reduced level of GENSET gene expression, the method comprising the steps of : a) providing a biological sample from the mammal; and b) comparing the amount of any

20   of the herein-described polypeptides, or of an RNA species encoding the polypeptide, within the biological sample with a level detected in or expected from a control sample; wherein an increased amount of the polypeptide or the RNA species within the biological sample compared to the level detected in or expected from the control sample indicates that the mammal has an elevated level of the GENSET gene expression, and wherein a decreased amount of the polypeptide or the RNA

25   species within the biological sample compared to the level detected in or expected from the control sample indicates that the mammal has a reduced level of the GENSET gene expression.

In another aspect, the present invention provides a method of identifying a candidate modulator of a GENSET polypeptide, the method comprising : a) contacting any of the herein-described polypeptides with a test compound; and b) determining whether the compound

30   specifically binds to the polypeptide; wherein a detection that the compound specifically binds to the polypeptide indicates that the compound is a candidate modulator of the GENSET polypeptide.

Brief description of drawings

Figure 1 is a map of the expression vector pPT

35   Figure 2 is a block diagram of an exemplary computer system.

13

Figure 3 is a flow diagram illustrating one embodiment of a process 200 for comparing a new nucleotide or protein sequence with a database of sequences in order to determine the identity levels between the new sequence and the sequences in the database.

Figure 4 is a flow diagram illustrating one embodiment of a process 250 in a computer for determining whether two sequences are homologous.

Figure 5 is a flow diagram illustrating one embodiment of an identifier process 300 for detecting the presence of a feature in a sequence.

Brief Description of Tables

Table I provides the applicant's internal designation number assigned to each sequence identification number and indicates whether the sequence is a nucleic acid sequence or a polypeptide sequence, and in which vector the cDNA was cloned.

Table II provides structural features for each cDNA of SEQ ID Nos: 1-241 i.e., the locations of the full coding sequences, the signal peptides, the mature polypeptides, the polyA signal and the polyA site.

Table III lists variants for cDNAs of the present invention.

Table IV provides the positions of fragments which are preferably excluded from the present invention.

Tables Va and b provides the positions of fragments which are preferably excluded or included in the present invention. Table IV and Tables Va, and Table Vb provide for the inclusion and exclusion of polynucleotides independently from each other in addition to those described elsewhere in the specification and is therefore, not meant as limiting description.

Table VI lists known biologically structural and functional domains for the polypeptides of the present invention.

Table VII lists antigenic peaks of predicted antigenic epitopes for polypeptides of the present invention.

Table VIII lists the putative chromosomal location of the polynucleotides of the present invention.

Table IX list the Genset's cDNA libraries of tissues and cell types examined that express the polynucleotides of the present invention.

Table X relates to the bias in spatial distribution of the polynucleotide sequences of the present invention.

Table XI lists predicted subcellular localization for cDNAs of the present invention.

Table XII gives the correspondence between the polynucleotides of the US priority applications, namely the US Provisional Patent Applications Serial Nos 60/169,629 and 60/187, (column entitled "Seq Id No in priority applications") and the polynucleotides of the present application (column entitled "Seq Id No in present application").

14

Brief description of sequence listing

SEQ ID Nos: 1-31 and 33-143 are the nucleotide sequences of cDNAs encoding a potentially secreted protein. The locations of the ORFs and sequences encoding signal peptides are listed in the accompanying Sequence Listing. In addition, the von Heijne score of the signal peptide

5   computed as described below is listed as the "score" in the accompanying Sequence Listing. The sequence of the signal-peptide is listed as "seq" in the accompanying Sequence Listing. The "/" in the signal peptide sequence indicates the location where proteolytic cleavage of the signal peptide occurs to generate a mature protein. When appropriate, the locations of the first and last nucleotides of the coding sequences, eventually the locations of the first and last nucleotides of the polyA and

10 the locations of the first and last nucleotides of the polyA sites are indicated.

SEQ ID Nos. 32 and 144-241 are the nucleotide sequences of cDNAs in which no sequence encoding a signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a sequence encoding a signal peptide in these nucleic acids. The locations of the ORFs are listed in the accompanying Sequence Listing. When appropriate, the locations of the

15 first and last nucleotides of the coding sequences, eventually the locations of the first and last nucleotides of the polyA and the locations of the first and last nucleotides of the polyA sites are indicated.

SEQ ID Nos: 242-272 and 274-384 are the amino acid sequences of polypeptides which contain a signal peptide. These polypeptides are encoded by the cDNAs of SEQ ID Nos: 1-31 and

20 33-143 respectively. The location of the signal peptide is listed in the accompanying Sequence Listing.

SEQ ID Nos: 273 and 385-482 are the amino acid sequences of polypeptides in which no signal peptide has been identified to date. However, it remains possible that subsequent analysis will identify a signal peptide in these polypeptides. These polypeptides are encoded by the nucleic

25 acids of SEQ ID Nos: 32 and 144-241 respectively.

In accordance with the regulations relating to Sequence Listings, the following codes have been used in the Sequence Listing to describes nucleotide sequences. The code "r" in the sequences indicates that the nucleotide may be a guanine or an adenine. The code "y" in the sequences indicates that the nucleotide may be a thymine or a cytosine. The code "m" in the sequences

30 indicates that the nucleotide may be an adenine or a cytosine. The code "k" in the sequences indicates that the nucleotide may be a guanine or a thymine. The code "s" in the sequences indicates that the nucleotide may be a guanine or a cytosine. The code "w" in the sequences indicates that the nucleotide may be an adenine or an thymine. In addition, all instances of the symbol "n" in the nucleic acid sequences mean that the nucleotide can be adenine, guanine, cytosine or thymine.

35 In some instances, the polypeptide sequences in the Sequence Listing contain the symbol "Xaa." These "Xaa" symbols indicate either (1) a residue which cannot be identified because of nucleotide sequence ambiguity or (2) a stop codon in the determined sequence where applicants

believe one should not exist (if the sequence were determined more accurately). In some instances, several possible identities of the unknown amino acids may be suggested by the genetic code.

In the case of secreted proteins, it should be noted that, in accordance with the regulations governing Sequence Listings, in the appended Sequence Listing, the encoded protein (i.e. the

5   protein containing the signal peptide and the mature protein or part thereof) extends from an amino acid residue having a negative number through a positively numbered amino acid residue. Thus, the first amino acid of the mature protein resulting from cleavage of the signal peptide is designated as amino acid number 1, and the first amino acid of the signal peptide is designated with the appropriate negative number. However, in the present application, positions on amino acid

10   sequences are always given on the full length polypeptide, the first amino acid of the signal peptide being designated as amino acid number 1.

Detailed description

DEFINITIONS

Before describing the invention in greater detail, the following definitions are set forth to

15   illustrate and define the meaning and scope of the terms used to describe the invention herein.

The terms "GENSET gene", when used herein, encompasses genomic, mRNA and cDNA sequences encoding the GENSET protein, including the 5' and 3' untranslated regions of said sequences.

As used herein, a "secreted" protein is one which, when expressed in a suitable host cell, is

20   transported across or through a membrane, including transport as a result of signal peptides in its amino acid sequence. "Secreted" proteins include without limitation proteins secreted wholly (e.g. soluble proteins), or partially (e.g. receptors) from the cell in which they are expressed. "Secreted" proteins also include without limitation proteins which are transported across the membrane of the endoplasmic reticulum. As used herein, a "mature protein" is the polypeptide fragment generated

25   after the cleavage of the signal peptide.

The term "full coding sequence" or open reading frame (ORF) of a GENSET gene, when used herein, refers to the complete coding sequence of said gene. In the case of a secreted protein, the full coding sequence comprises the coding sequence for the signal peptide and the coding sequence for the mature polypeptide. Accordingly, the term "full-length polypeptide" refers to the

30   complete polypeptide encoded by said GENSET gene and in the case of a secreted protein it comprises both the signal peptide and the mature polypeptide. The positions of the full length polypeptides and, in the case of secreted proteins, of signal peptides and mature polypeptides are given in the appended sequence listing.

The term "GENSET biological activity" is intended for polypeptides exhibiting an activity

35   similar, but not necessarily identical, to an activity of the GENSET polypeptide of the invention.

16

The GENSET biological activity of a given polypeptide may be assessed using a suitable biological assay well known to those skilled in the art such as the one(s) described herein. In contrast, the term "biological activity" refers to any activity that a polypeptide of the invention may have.

The term "corresponding mRNA" refers to the mRNA which was the template for the

5   cDNA synthesis which produced a cDNA of the present invention.

The term "corresponding genomic DNA" refers to the genomic DNA which encodes mRNA which includes the sequence of one of the strands of the cDNA in which thymidine residues in the sequence of the cDNA are replaced by uracil residues in the mRNA.

The term "deposited clone pool" is used herein to refer to the pool of clones entitled

10   GENSET.071PRF deposited in ATCC with the accession number PTA-1218 on January, 21, 2000.

The term "heterologous", when used herein, is intended to designate any polynucleotide or polypeptide other than the GENSET polynucleotide or polypeptide respectively.

The term "isolated" requires that the material be removed from its original environment (e. g., the natural environment if it is naturally occurring). For example, a naturally-occurring

15   polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or DNA or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotide could be part of a vector and/or such polynucleotide or polypeptide could be part of a composition, and still be isolated in that the vector or composition is not part of its natural environment. For example, a naturally-occurring polynucleotide present in a living

20   animal is not isolated, but the same polynucleotide, separated from some or all of the coexisting materials in the natural system, is isolated. Specifically excluded from the definition of "isolated" are: naturally-occurring chromosomes (such as chromosome spreads), artificial chromosome libraries, genomic libraries, and cDNA libraries that exist either as an *in vitro* nucleic acid preparation or as a transfected/transformed host cell preparation, wherein the host cells are either an

25   *in vitro* heterogeneous preparation or plated as a heterogeneous population of single colonies. Also specifically excluded are the above libraries wherein a specified polynucleotide makes up less than 5% of the number of nucleic acid inserts in the vector molecules. Further specifically excluded are whole cell genomic DNA or whole cell RNA preparations (including said whole cell preparations which are mechanically sheared or enzymatically digested). Further specifically excluded are the

30   above whole cell preparations as either an *in vitro* preparation or as a heterogeneous mixture separated by electrophoresis (including blot transfers of the same) wherein the polynucleotide of the invention has not further been separated from the heterologous polynucleotides in the electrophoresis medium (e.g., further separating by excising a single band from a heterogeneous band population in an agarose gel or nylon blot).

35   The term "purified" does not require absolute purity; rather, it is intended as a relative definition. Purification of starting material or natural material to at least one order of magnitude, preferably two or three orders, and more preferably four or five orders of magnitude is expressly

17

contemplated. As an example, purification from 0.1 % concentration to 10 % concentration is two orders of magnitude. To illustrate, individual cDNA clones isolated from a cDNA library have been conventionally purified to electrophoretic homogeneity. The sequences obtained from these clones could not be obtained directly either from the library or from total human DNA. The cDNA clones

5 are not naturally occurring as such, but rather are obtained via manipulation of a partially purified naturally occurring substance (messenger RNA). The conversion of mRNA into a cDNA library involves the creation of a synthetic substance (cDNA) and pure individual cDNA clones can be isolated from the synthetic library by clonal selection. Thus, creating a cDNA library from messenger RNA and subsequently isolating individual clones from that library results in an

10 approximately 10^4-10^6 fold purification of the native message.

     The term "purified" is further used herein to describe a polypeptide or polynucleotide of the invention which has been separated from other compounds including, but not limited to, polypeptides or polynucleotides, carbohydrates, lipids, etc. The term "purified" may be used to specify the separation of monomeric polypeptides of the invention from oligomeric forms such as

15 homo- or hetero- dimers, trimers, etc. The term "purified" may also be used to specify the separation of covalently closed polynucleotides from linear polynucleotides. A polynucleotide is substantially pure when at least about 50%, preferably 60 to 75% of a sample exhibits a single polynucleotide sequence and conformation (linear versus covalently close). A substantially pure polypeptide or polynucleotide typically comprises about 50%, preferably 60 to 90% weight/weight

20 of a polypeptide or polynucleotide sample, respectively, more usually about 95%, and preferably is over about 99% pure. Polypeptide and polynucleotide purity, or homogeneity, is indicated by a number of means well known in the art, such as agarose or polyacrylamide gel electrophoresis of a sample, followed by visualizing a single band upon staining the gel. For certain purposes higher resolution can be provided by using HPLC or other means well known in the art. As an alternative

25 embodiment, purification of the polypeptides and polynucleotides of the present invention may be expressed as "at least" a percent purity relative to heterologous polypeptides and polynucleotides (DNA, RNA or both). As a preferred embodiment, the polypeptides and polynucleotides of the present invention are at least; 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 96%, 96%, 98%, 99%, or 100% pure relative to heterologous polypeptides and polynucleotides, respectively. As a

30 further preferred embodiment the polypeptides and polynucleotides have a purity ranging from any number, to the thousandth position, between 90% and 100% (e.g., a polypeptide or polynucleotide at least 99.995% pure) relative to either heterologous polypeptides or polynucleotides, respectively, or as a weight/weight ratio relative to all compounds and molecules other than those existing in the carrier. Each number representing a percent purity, to the thousandth position, may be claimed as

35 individual species of purity.

     As used interchangeably herein, the terms "nucleic acid molecule(s)", "oligonucleotide(s)", and "polynucleotide(s)" include RNA or DNA (either single or double stranded, coding,

complementary or antisense), or RNA/DNA hybrid sequences of more than one nucleotide in either single chain or duplex form (although each of the above species may be particularly specified). The term "nucleotide" is used herein as an adjective to describe molecules comprising RNA, DNA, or RNA/DNA hybrid sequences of any length in single-stranded or duplex form. More precisely, the

5   expression "nucleotide sequence" encompasses the nucleic material itself and is thus not restricted to the sequence information (i.e. the succession of letters chosen among the four base letters) that biochemically characterizes a specific DNA or RNA molecule. The term "nucleotide" is also used herein as a noun to refer to individual nucleotides or varieties of nucleotides, meaning a molecule, or individual unit in a larger nucleic acid molecule, comprising a purine or pyrimidine, a ribose or

10   deoxyribose sugar moiety, and a phosphate group, or phosphodiester linkage in the case of nucleotides within an oligonucleotide or polynucleotide. The term "nucleotide" is also used herein to encompass "modified nucleotides" which comprise at least one modifications such as (a) an alternative linking group, (b) an analogous form of purine, (c) an analogous form of pyrimidine, or (d) an analogous sugar. For examples of analogous linking groups, purine, pyrimidines, and sugars

15   see for example PCT publication No. WO 95/04064, which disclosure is hereby incorporated by reference in its entirety. Preferred modifications of the present invention include, but are not limited to, 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xantine, 4-acetylcytosine, 5-(carboxyhydroxylmethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N6-

20   isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methylcytosine, N6-adenine, 7-methylguanine, 5-methylaminomethyluracil, 5-methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N6-isopentenyladenine, uracil-5-oxyacetic acid (v) ybutoxosine, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-

25   thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-5-oxyacetic acid, 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, and 2,6-diaminopurine. The polynucleotide sequences of the invention may be prepared by any known method, including synthetic, recombinant, *ex vivo* generation, or a combination thereof, as well as utilizing any purification methods known in the art. Methylenemethylimino linked oligonucleosides as well as

30   mixed backbone compounds having, may be prepared as described in U.S. Pat. Nos. 5,378,825; 5,386,023; 5,489,677; 5,602,240; and 5,610,289, which disclosures are hereby incorporated by reference in their entireties. Formacetal and thioformacetal linked oligonucleosides may be prepared as described in U.S. Pat. Nos. 5,264,562 and 5,264,564, which disclosures are hereby incorporated by reference in their entireties. Ethylene oxide linked oligonucleosides may be prepared as

35   described in U.S. Pat. No. 5,223,618, which disclosure is hereby incorporated by reference in its entirety. Phosphinate oligonucleotides may be prepared as described in U.S. Pat. No. 5,508,270, which disclosure is hereby incorporated by reference in its entirety. Alkyl phosphonate

19

oligonucleotides may be prepared as described in U.S. Pat. No. 4,469,863, which disclosure is hereby incorporated by reference in its entirety. 3'-Deoxy-3'-methylene phosphonate oligonucleotides may be prepared as described in U.S. Pat. Nos. 5,610,289 or 5,625,050 which disclosures are hereby incorporated by reference in their entireties. Phosphoramidite

5   oligonucleotides may be prepared as described in U.S. Pat. No. 5,256,775 or U.S. Pat. No. 5,366,878 which disclosures are hereby incorporated by reference in their entireties. Alkylphosphonothioate oligonucleotides may be prepared as described in published PCT applications WO 94/17093 and WO 94/02499 which disclosures are hereby incorporated by reference in their entireties. 3'-Deoxy-3'-amino phosphoramidate oligonucleotides may be prepared

10   as described in U.S. Pat. No. 5,476,925, which disclosure is hereby incorporated by reference in its entirety. Phosphotriester oligonucleotides may be prepared as described in U.S. Pat. No. 5,023,243, which disclosure is hereby incorporated by reference in its entirety. Borano phosphate oligonucleotides may be prepared as described in U.S. Pat. Nos. 5,130,302 and 5,177,198 which disclosures are hereby incorporated by reference in their entireties.

15   The term "upstream" is used herein to refer to a location which is toward the 5' end of the polynucleotide from a specific reference point.

The terms "base paired" and "Watson & Crick base paired" are used interchangeably herein to refer to nucleotides which can be hydrogen bonded to one another be virtue of their sequence identities in a manner like that found in double-helical DNA with thymine or uracil residues linked

20   to adenine residues by two hydrogen bonds and cytosine and guanine residues linked by three hydrogen bonds (See Stryer, 1995, which disclosure is hereby incorporated by reference in its entirety).

The terms "complementary" or "complement thereof" are used herein to refer to the sequences of polynucleotides which is capable of forming Watson & Crick base pairing with

25   another specified polynucleotide throughout the entirety of the complementary region. For the purpose of the present invention, a first polynucleotide is deemed to be complementary to a second polynucleotide when each base in the first polynucleotide is paired with its complementary base. Complementary bases are, generally, A and T (or A and U), or C and G. "Complement" is used herein as a synonym from "complementary polynucleotide", "complementary nucleic acid" and

30   "complementary nucleotide sequence". These terms are applied to pairs of polynucleotides based solely upon their sequences and not any particular set of conditions under which the two polynucleotides would actually bind. Unless otherwise stated, all complementary polynucleotides are fully complementary on the whole length of the considered polynucleotide.

The terms "polypeptide" and "protein", used interchangeably herein, refer to a polymer of

35   amino acids without regard to the length of the polymer; thus, peptides, oligopeptides, and proteins are included within the definition of polypeptide. This term also does not specify or exclude chemical or post-expression modifications of the polypeptides of the invention, although chemical

or post-expression modifications of these polypeptides may be included excluded as specific embodiments. Therefore, for example, modifications to polypeptides that include the covalent attachment of glycosyl groups, acetyl groups, phosphate groups, lipid groups and the like are expressly encompassed by the term polypeptide. Further, polypeptides with these modifications

5   may be specified as individual species to be included or excluded from the present invention. The natural or other chemical modifications, such as those listed in examples above can occur anywhere in a polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given polypeptide. Also, a given polypeptide may

10   contain many types of modifications. Polypeptides may be branched, for example, as a result of ubiquitination, and they may be cyclic, with or without branching. Modifications include acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of phosphotidylinositol, cross-linking,

15   cyclization, disulfide bond formation, demethylation, formation of covalent cross-links, formation of cysteine, formation of pyroglutamate, formylation, gamma-carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristoylation, oxidation, pegylation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, transfer-RNA mediated addition of amino acids to proteins such as arginylation, and ubiquitination.

20   (See, for instance Creighton (1993); Seifter *et al.*, (1990); Rattan *et al.*, (1992)). Also included within the definition are polypeptides which contain one or more analogs of an amino acid (including, for example, non-naturally occurring amino acids, amino acids which only occur naturally in an unrelated biological system, modified amino acids from mammalian systems, etc...), polypeptides with substituted linkages, as well as other modifications known in the art, both

25   naturally occurring and non-naturally occurring.

As used herein, the terms "recombinant polynucleotide" and "polynucleotide construct" are used interchangeably to refer to linear or circular, purified or isolated polynucleotides that have been artificially designed and which comprise at least two nucleotide sequences that are not found as contiguous nucleotide sequences in their initial natural environment. In particular, this terms

30   mean that the polynucleotide or cDNA is adjacent to "backbone" nucleic acid to which it is not adjacent in its natural environment. Additionally, to be "enriched" the cDNAs will represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid backbone molecules. Backbone molecules according to the present invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic

35   acids used to maintain or manipulate a nucleic acid insert of interest. Preferably, the enriched cDNAs represent 15% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. More preferably, the enriched cDNAs represent 50% or more of

the number of nucleic acid inserts in the population of recombinant backbone molecules. In a highly preferred embodiment, the enriched cDNAs represent 90% or more (including any number between 90 and 100%, to the thousandth position, e.g., 99.5%) # of the number of nucleic acid inserts in the population of recombinant backbone molecules.

5   The term "recombinant polypeptide" is used herein to refer to polypeptides that have been artificially designed and which comprise at least two polypeptide sequences that are not found as contiguous polypeptide sequences in their initial natural environment, or to refer to polypeptides which have been expressed from a recombinant polynucleotide.

        As used herein, the term "operably linked" refers to a linkage of polynucleotide elements in 10 a functional relationship. A sequence which is "operably linked" to a regulatory sequence such as a promoter means that said regulatory element is in the correct location and orientation in relation to the nucleic acid to control RNA polymerase initiation and expression of the nucleic acid of interest. For instance, a promoter or enhancer is operably linked to a coding sequence if it affects the transcription of the coding sequence.

15   As used herein, the term "non-human animal" refers to any non-human animal, including insects, birds, rodents and more usually mammals. Preferred non-human animals include: primates; farm animals such as swine, goats, sheep, donkeys, cattle, horses, chickens, rabbits; and rodents, preferably rats or mice. As used herein, the term "animal" is used to refer to any species in the animal kingdom, preferably vertebrates, including birds and fish, and more preferable a mammal. 20 Both the terms "animal" and "mammal" expressly embrace human subjects unless preceded with the term "non-human".

        The term "domain" refers to an amino acid fragment with specific biological properties. This term encompasses all known structural and linear biological motifs. Examples of such motifs include but are not limited to leucine zippers, helix-turn-helix motifs, glycosylation sites, 25 ubiquitination sites, alpha helices, and beta sheets, signal peptides which direct the secretion of proteins, sites for post-translational modification, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

        Although they have distinct meanings, the terms "comprising", "consisting of" and "consisting essentially of" may be interchanged for one another throughout the instant application". 30 The term "having" has the same meaning as "comprising" and may be replaced with either the term "consisting of" or "consisting essentially of".

        An "amplification product" refers to a product of any amplification reaction, e.g. PCR, RT-PCR, LCR, etc.

        A "modulator" of a protein or other compound refers to any agent that has a functional 35 effect on the protein, including physical binding to the protein, alterations of the quantity or quality of expression of the protein, altering any measurable or detectable activity, property, or behavior of the protein, or in any way interacts with the protein or compound.

"A test compound" can be any molecule that is evaluated for its ability to modulate a protein or other compound.

Unless otherwise specified in the application, nucleotides and amino acids of polynucleotides and polypeptides respectively of the present invention are contiguous and not
5   interrupted by heterologous sequences.

Identity Between Nucleic Acids Or Polypeptides

The terms "percentage of sequence identity" and "percentage homology" are used interchangeably herein to refer to comparisons among polynucleotides and polypeptides, and are determined by comparing two optimally aligned sequences over a comparison window, wherein the
10 portion of the polynucleotide or polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which the identical nucleic acid base or amino acid residue occurs in both sequences to yield the number of matched positions, dividing the number of matched
15 positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence identity. Homology is evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are by no means limited to, TBLASTN, BLASTP, FASTA, TFASTA, CLUSTALW, FASTDB (Pearson and Lipman, 1988; Altschul *et al.*, 1990; Thompson *et al.*, 1994; Higgins *et*
20 *al.*, 1996; Altschul *et al.*, 1990; Altschul *et al.*, 1993; Brutlag *et al*, 1990), the disclosures of which are incorporated by reference in their entireties.

In a particularly preferred embodiment, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST") which is well known in the art (see, e.g., Karlin and Altschul, 1990; Altschul *et al.*, 1990, 1993, 1997), the disclosures of which
25 are incorporated by reference in their entireties. In particular, five specific BLAST programs are used to perform the following task:

(1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database;

(2) BLASTN compares a nucleotide query sequence against a nucleotide sequence
30 database;

(3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database;

(4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and
35 (5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring

5   matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet et al., 1992; Henikoff and Henikoff, 1993), the disclosures of which are incorporated by reference in their entireties. Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978), the disclosure of which is incorporated by reference in its entirety. The BLAST programs evaluate the statistical significance of all high-

10 scoring segment pairs identified, and preferably selects those segments which satisfy a user-specified threshold of significance, such as a user-specified percent homology. Preferably, the statistical significance of a high-scoring segment pair is evaluated using the statistical significance formula of Karlin (see, e.g., Karlin and Altschul, 1990), the disclosure of which is incorporated by reference in its entirety. The BLAST programs may be used with the default parameters or with

15 modified parameters provided by the user.

Another preferred method for determining the best overall match between a query nucleotide sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag et al. (1990), the disclosure of which is incorporated by reference in its

20 entirety. In a sequence alignment the query and subject sequences are both DNA sequences. An RNA sequence can be compared by first converting U's to T's. The result of said global sequence alignment is in percent identity. Preferred parameters used in a FASTDB alignment of DNA sequences to calculate percent identity are: Matrix=Unitary, k-tuple=4, Mismatch Penalty= 1, Joining Penalty=30, Randomization Group Length=0, Cutoff Score= 1, Gap Penalty=5, Gap Size

25 Penalty 0.05, Window Size=500 or the length of the subject nucleotide sequence, whichever is 35 shorter. If the subject sequence is shorter than the query sequence because of 5' or 3' deletions, not because of internal deletions, a manual correction must be made to the results. This is because the FASTDB program does not account for 5' and 3' truncations of the subject sequence when calculating percent identity. For subject sequences truncated at the 5' or 3'ends, relative to the query

30 sequence, the percent identity is corrected by calculating the number of bases of the query sequence that are 5' and 3' of the subject sequence, which are not matched/aligned, as a percent of the total bases of the query sequence. Whether a nucleotide is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using 10, the specified parameters, to arrive at a final

35 percent identity score. This corrected score is what is used for the purposes of the present invention. Only nucleotides outside the 5' and 3' nucleotides of the subject sequence, as displayed by the FASTDB alignment, which are not matched/aligned with the query sequence, are calculated for the

purposes of manually adjusting the percent identity score. For example, a 90 nucleotide subject sequence is aligned to a 100 nucleotide query sequence to determine percent identity. The deletions occur at the 5' end of the subject sequence and therefore, the FASTDB alignment does not show a matched/alignment of the first 10 nucleotides at 5' end. The 10 unpaired nucleotides represent 10%

5 of the sequence (number of nucleotides at the 5' and 3' ends not matched/total number of nucleotides in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 nucleotides were perfectly matched the final percent identity would be 90%. In another example, a 90 nucleotide subject sequence is compared with a 100 nucleotide query sequence. This time the deletions are internal deletions so that there are no

10 nucleotides on the 5' or 3' of the subject sequence which are not matched/aligned with the query. In this case the percent identity calculated by FASTDB is not manually corrected. Once again, only nucleotides 5' and 3' of the subject sequence which are not matched/aligned with the query sequence are manually corrected. No other manual corrections are made for the purposes of the present invention.

15   Another preferred method for determining the best overall match between a query amino acid sequence (a sequence of the present invention) and a subject sequence, also referred to as a global sequence alignment, can be determined using the FASTDB computer program based on the algorithm of Brutlag et al. (1990). In a sequence alignment the query and subject sequences are both amino acid sequences. The result of said global sequence alignment is in percent identity. Preferred

20 parameters used in a FASTDB amino acid alignment are: Matrix=PAM 0, k-tuple=2, Mismatch Penalty= 1, Joining Penalty=20, Randomization Group25Length=0, Cutoff Score= 1, Window Size=sequence length, Gap Penalty=5, Gap Size Penalty=0.05, Window Size=500 or the length of the subject amino acid sequence, whichever is shorter. If the subject sequence is shorter than the query sequence due to N-or C-terminal deletions, not because of internal deletions, the results, in

25 percent identity, must be manually corrected. This is because the FASTDB program does not account for N- and C-terminal truncations of the subject sequence when calculating global percent identity. For subject sequences truncated at the N- and C-termini, relative to the query sequence, the percent identity is corrected by calculating the number of residues of the query sequence that are N- and C- terminal of the subject sequence, which are not matched/aligned with a corresponding

30 subject residue, as a percent of the total bases of the query sequence. Whether a residue is matched/aligned is determined by results of the FASTDB sequence alignment. This percentage is then subtracted from the percent identity, calculated by the above FASTDB program using the specified parameters, to arrive at a final percent identity score. This final percent identity score is what is used for the purposes of the present invention. Only residues to the N- and C-termini of the

35 subject sequence, which are not matched/aligned with the query sequence, are considered for the purposes of manually adjusting the percent identity score. That is, only query amino acid residues outside the farthest N- and C-terminal residues of the subject sequence. For example, a 90 amino

acid residue subject sequence is aligned with a 100-residue query sequence to determine percent identity. The deletion occurs at the N-terminus of the subject sequence and therefore, the FASTDB alignment does not match/align with the first residues at the N-terminus. The 10 unpaired residues represent 10% of the sequence (number of residues at the N- and C- termini not matched/total

5   number of residues in the query sequence) so 10% is subtracted from the percent identity score calculated by the FASTDB program. If the remaining 90 residues were perfectly matched the final percent identity would be 90%. In another example, a 90-residue subject sequence is compared with a 100-residue query sequence. This time the deletions are internal so there are no residues at the N- or C-termini of the subject sequence, which are not matched/aligned with the query. In this case the

10   percent identity calculated by FASTDB is not manually corrected. Once again, only residue positions outside the N- and C-terminal ends of the subject sequence, as displayed in the FASTDB alignment, which are not matched/aligned with the query sequence are manually corrected. No other manual corrections are made for the purposes of the present invention.

The term "percentage of sequence similarity" refers to comparisons between polypeptide
15   sequences and is determined by comparing two optimally aligned sequences over a comparison window, wherein the portion of the polypeptide sequence in the comparison window may comprise additions or deletions (i.e., gaps) as compared to the reference sequence (which does not comprise additions or deletions) for optimal alignment of the two sequences. The percentage is calculated by determining the number of positions at which an identical or equivalent amino acid residue occurs

20   in both sequences to yield the number of matched positions, dividing the number of matched positions by the total number of positions in the window of comparison and multiplying the result by 100 to yield the percentage of sequence similarity. Similarity is evaluated using any of the variety of sequence comparison algorithms and programs known in the art, including those described above in this section. Equivalent amino acid residues are defined herein in the "Mutated

25   polypeptides" section.

POLYNUCLEOTIDES OF THE INVENTION

The present invention concerns GENSET genomic and cDNA sequences. The present invention encompasses GENSET genes, polynucleotides comprising GENSET genomic and cDNA sequences, as well as fragments and variants thereof. These polynucleotides may be purified,
30   isolated, or recombinant.

Also encompassed by the present invention are allelic variants, orthologs, splice variants, and/or species homologues of the GENSET genes. Procedures known in the art can be used to obtain full-length genes and cDNAs, allelic variants, splice variants, full-length coding portions, orthologs, and/or species homologues of genes and cDNAs corresponding to a nucleotide sequence
35   selected from the group consisting of sequences of SEQ ID Nos: 1-241 and sequences of clone inserts of the deposited clone pool, using information from the sequences disclosed herein or the

26

clone pool deposited with the ATCC. For example, allelic variants, orthologs and/or species
homologues may be isolated and identified by making suitable probes or primers from the
sequences provided herein and screening a suitable nucleic acid source for allelic variants and/or the
desired homologue using any technique known to those skilled in the art including those described

5 into the section entitled "To find similar sequences".

In a specific embodiment, the polynucleotides of the invention are at least 15, 30, 50, 100,
125, 500, or 1000 continuous nucleotides. In another embodiment, the polynucleotides are less than
or equal to 300kb, 200kb, 100kb, 50kb, 10kb, 7.5kb, 5kb, 2.5kb, 2kb, 1.5kb, or 1kb in length. In a
further embodiment, polynucleotides of the invention comprise a portion of the coding sequences,

10 as disclosed herein, but do not comprise all or a portion of any intron. In another embodiment, the
polynucleotides comprising coding sequences do not contain coding sequences of a genomic
flanking gene (i.e., 5' or 3' to the gene of interest in the genome). In other embodiments, the
polynucleotides of the invention do not contain the coding sequence of more than 1000, 500, 250,
100, 75, 50, 25, 20, 15, 10, 5, 4, 3, 2, or 1 naturally occurring genomic flanking gene(s).

15 Deposited clone pool of the invention

Expression of GENSET genes has been shown to lead to the production of at least one
mRNA species per GENSET gene, which cDNA sequence is set forth in the appended sequence
listing as SEQ ID Nos: 1-241. The cDNAs (SEQ ID Nos: 1-241) corresponding to these GENSET
mRNA species were cloned in the vector pBluescriptII SK⁻ (Stratagene) or one of its derivative

20 called pPT (see figure 1). Cells containing the cloned cDNAs of the present invention are
maintained in permanent deposit by the inventors at Genset, S.A., 24 Rue Royale, 75008 Paris,
France. Table I provides the applicant's internal designation number (column entitled "Internal
designation") assigned to each sequence identification number of SEQ ID Nos: 1-482 (column
entitled "Seq Id No") and indicates whether the sequence is a nucleic acid sequence or a

25 polypeptide sequence (column entitled "Type"), and in which vector the cDNA was cloned (column
entitled "Vector").

Each cDNA can be removed from the Bluescript vector in which it was inserted by
performing a NotI Pst I double digestion to produce the appropriate fragment for each clone
provided the cDNA sequence of interest does not contain this restriction site within its sequence.

30 The preferable sites for cDNA removal for those clones inserted into pPT are MunI and HindIII, the
sites used for cloning provided the cDNA sequence of interest does not contain this restriction site
within its sequence. Alternatively, other restriction enzymes of the multicloning site of the vector
may be used to recover the desired insert as indicated by the manufacturer or in figure 1.

Pool of cells containing the cDNAs of the invention, from which the cells containing a

35 particular polynucleotide is obtainable, were also deposited with the American Tissue Culture
Collection (ATCC), 10801 University Boulevard, Manassas, VA 20110-2209, United States . Each

cDNA clone has been transfected into separate bacterial cells (E-coli) for these composite deposits. In particular, cells containing the sequences of SEQ ID Nos: 1-241 were deposited on January, 21, 2000 in the pool having ATCC Accession No. PTA-1218 and designated GENSET.071PRF.

Bacterial cells containing a particular clone can be obtained from the composite deposit as
5   follows:

An oligonucleotide probe or probes should be designed to the sequence that is known for that particular clone. This sequence can be derived from the sequences provided herein, or from a combination of those sequences. The design of the oligonucleotide probe should preferably follow these parameters:

10   (a) It should be designed to an area of the sequence which has the fewest ambiguous bases ("N's"), if any;

(b) Preferably, the probe is designed to have a Tm of approximately 80 degree Celsius (assuming 2 degrees for each A or T and 4 degrees for each G or C). However, probes having melting temperatures between 40 degree Celsius and 80 degree Celsius may also be used provided
15 that specificity is not lost.

The oligonucleotide should preferably be labeled with gamma[^{32}P]ATP (specific activity 6000 Ci/mmole) and T4 polynucleotide kinase using commonly employed techniques for labeling oligonucleotides. Other labeling techniques can also be used. Unincorporated label should preferably be removed by gel filtration chromatography or other established methods. The amount
20 of radioactivity incorporated into the probe should be quantified by measurement in a scintillation counter. Preferably, specific activity of the resulting probe should be approximately 4x10^6 dpm/pmole.

The bacterial culture containing the pool of full-length clones should preferably be thawed and 100 ul of the stock used to inoculate a sterile culture flask containing 25 ml of sterile L-broth
25 containing ampicillin at 100 ug/ml. The culture should preferably be grown to saturation at 37 degree Celsius, and the saturated culture should preferably be diluted in fresh L-broth. Aliquots of these dilutions should preferably be plated to determine the dilution and volume which will yield approximately 5000 distinct and well-separated colonies on solid bacteriological media containing L-broth containing ampicillin at 100 ug/ml and agar at 1.5% in a 150 mm petri dish when grown
30 overnight at 37 degree Celsius. Other known methods of obtaining distinct, well-separated colonies can also be employed.

Standard colony hybridization procedures should then be used to transfer the colonies to nitrocellulose filters and lyse, denature and bake them.

The filter is then preferably incubated at 65 degree Celsius for 1 hour with gentle agitation
35 in 6X SSC (20X stock is 175.3 g NaCl/liter, 88.2 g Na citrate/liter, adjusted to pH 7.0 with NaOH) containing 0.5% SDS, 100 pg/ml of yeast RNA, and 10 mM EDTA (approximately 10 ml per 150 mm filter). Preferably, the probe is then added to the hybridization mix at a concentration greater

28

than or equal to 1x10⁶ dpm/ml. The filter is then preferably incubated at 65 degree Celsius with gentle agitation overnight. The filter is then preferably washed in 500 ml of 2X SSC/0.1% SDS at room temperature with gentle shaking for 15 minutes. A third wash with 0.1X SSC/0.5% SDS at 65 degree Celsius for 30 minutes to 1 hour is optional. The filter is then preferably dried and subjected

5   to autoradiography for sufficient time to visualize the positives on the X-ray film. Other known hybridization methods can also be employed.

The positive colonies are picked, grown in culture, and plasmid DNA isolated using standard procedures. The clones can then be verified by restriction analysis, hybridization analysis, or DNA sequencing. The plasmid DNA obtained using these procedures may then be manipulated

10   using standard cloning techniques familiar to those skilled in the art.

Alternatively, to recover cDNA inserts from the pool of bacteria, a PCR can be performed on plasmid DNA isolated using standard procedures and primers designed at both ends of the cDNA insertion, including primers designed in the multicloning site of the vector . For example, a PCR reaction may be conducted using universal primers designed by the plasmid provider or using

15   primers which are specific to the cDNA of interest In the case of Bluescript SK(-), a PCR reaction may be conducted using a primer having the sequence GGAAACAGCTATGACCA and a primer having the sequence GTAAAACGACGGCCAGT. This will produce a DNA fragment including a piece of the multiple cloning site and the cDNA insert. If a specific cDNA of interest is to be recovered, primers may be designed in order to be specific for the 5' end and the 3' end of this

20   cDNA using sequence information available from the appended sequence listing. The PCR product which corresponds to the cDNA of interest can then be manipulated using standard cloning techniques familiar to those skilled in the art.

Therefore, an object of the invention is an isolated, purified, or recombinant polynucleotide comprising a nucleotide sequence selected from the group consisting of cDNA inserts of the

25   deposited clone pool. Moreover, preferred polynucleotides of the invention include purified, isolated, or recombinant GENSET cDNAs consisting of, consisting essentially of, or comprising a nucleotide sequence selected from the group consisting of cDNA inserts of the deposited clone pool.

The polynucleotides of SEQ ID NOs: 1-141 may be interchanged with the corresponding

30   polynucleotides encoded by the human cDNA of the clones inserts of the deposited clone pool. The polypeptides of SEQ ID NOs: 242-482 may be interchanged with the corresponding polypeptides encoded by the human cDNA of the clones inserts of the deposited clone pool. The correspondance between the polynucleotides of SEQ ID Nos: 1-141, the polypeptides of SEQ ID NOs: 242-482 and clones inserts of the deposited clone pool is given in Table I..

invention. The following list is intended not to limit the invention but to provide guidance to programs and databases which are useful with the nucleic acid codes of the invention or the polypeptide codes of the invention. The programs and databases which may be used include, but are not limited to: MacPattern (EMBL), DiscoveryBase (Molecular Applications Group), GeneMine

5 (Molecular Applications Group), Look (Molecular Applications Group), MacLook (Molecular Applications Group), BLAST and BLAST2 (NCBI), BLASTN and BLASTX (Altschul *et al*, 1990), FASTA (Pearson and Lipman, 1988), FASTDB (Brutlag *et al.*, 1990), Catalyst (Molecular Simulations Inc.), Catalyst/SHAPE (Molecular Simulations Inc.), Cerius2.DBAccess (Molecular Simulations Inc.), HypoGen (Molecular Simulations Inc.), Insight II, (Molecular Simulations Inc.),

10 Discover (Molecular Simulations Inc.), CHARMm (Molecular Simulations Inc.), Felix (Molecular Simulations Inc.), DelPhi, (Molecular Simulations Inc.), QuanteMM, (Molecular Simulations Inc.), Homology (Molecular Simulations Inc.), Modeler (Molecular Simulations Inc.), ISIS (Molecular Simulations Inc.), Quanta/Protein Design (Molecular Simulations Inc.), WebLab (Molecular Simulations Inc.), WebLab Diversity Explorer (Molecular Simulations Inc.), Gene Explorer

15 (Molecular Simulations Inc.), SeqFold (Molecular Simulations Inc.), the EMBL/Swissprotein database, the MDL Available Chemicals Directory database, the MDL Drug Data Report data base, the Comprehensive Medicinal Chemistry database, Derwents's World Drug Index database, the BioByteMasterFile database, the Genbank database, and the Genseqn database. Many other programs and data bases would be apparent to one of skill in the art given the present disclosure.

20 Motifs which may be detected using the above programs include sequences encoding leucine zippers, helix-turn-helix motifs, glycosylation sites, ubiquitination sites, alpha helices, and beta sheets, signal sequences encoding signal peptides which direct the secretion of the encoded proteins, sequences implicated in transcription regulation such as homeoboxes, acidic stretches, enzymatic active sites, substrate binding sites, and enzymatic cleavage sites.

25 *CONCLUSION*

As discussed above, the GENSET polynucleotides and polypeptides of the present invention or fragments thereof can be used for various purposes. The polynucleotides can be used to express recombinant protein for analysis, characterization or therapeutic use; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a

30 particular stage of tissue differentiation or development or in disease states); as molecular weight markers on Southern gels; as chromosome markers or tags (when labeled) to identify chromosomes or to map related gene positions; as a reagent (including a labeled reagent) in assays designed to quantitatively determine levels of GENSET expression in biological samples; to compare with endogenous DNA sequences in patients to identify potential genetic disorders; as probes to

35 hybridize and thus discover novel, related DNA sequences; as a source of information to derive PCR primers for genetic fingerprinting; for selecting and making oligomers for attachment to a

450

"gene chip" or other support, including for examination for expression patterns; to raise anti-protein antibodies using DNA immunization techniques; and as an antigen to raise anti-DNA antibodies or elicit another immune response. Where the polynucleotide encodes a protein which binds or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the

5   polynucleotide can also be used in interaction trap assays (such as, for example, that described in Gyuris *et al.*, (1993) to identify polynucleotides encoding the other protein with which binding occurs or to identify inhibitors of the binding interaction.

The proteins or polypeptides provided by the present invention can similarly be used in assays to determine biological activity, including in a panel of multiple proteins for high-throughput

10   screening; to raise antibodies or to elicit another immune response; as a reagent (including the labeled reagent) in assays designed to quantitatively determine levels of the protein (or its receptor) in biological fluids; as markers for tissues in which the corresponding protein is preferentially expressed (either constitutively or at a particular stage of tissue differentiation or development or in a disease state); and, of course, to isolate correlative receptors or ligands. Where the protein binds

15   or potentially binds to another protein (such as, for example, in a receptor-ligand interaction), the protein can be used to identify the other protein with which binding occurs or to identify inhibitors of the binding interaction. Proteins involved in these binding interactions can also be used to screen for peptide or small molecule inhibitors or agonists of the binding interaction.

Any or all of these research utilities are capable of being developed into reagent grade or kit

20   format for commercialization as research products.

Methods for performing the uses listed above are well known to those skilled in the art. References disclosing such methods include without limitation "Molecular Cloning; A Laboratory Manual", 2d ed., Cole Spring Harbor Laboratory Press, Sambrook, J., E.F. Fritsch and T. Maniatis eds., 1989, and "Methods in Enzymology; Guide to Molecular Cloning Techniques", Academic

25   Press, Berger and Kimmel eds., 1987, which disclosures are hereby incorporated by reference in their entireties.

Polynucleotides and proteins of the present invention can also be used as nutritional sources or supplements. Such uses include without limitation use as a protein or amino acid supplement, use as a carbon source, use as a nitrogen source and use as a source of carbohydrate. In such cases

30   the protein or polynucleotide of the invention can be added to the feed of a particular organism or can be administered as a separate solid or liquid preparation, such as in the form of powder, pills, solutions, suspensions or capsules. In the case of microorganisms, the protein or polynucleotide of the invention can be added to the medium in or on which the microorganism is cultured.

Although this invention has been described in terms of certain preferred embodiments, other

35   embodiments which will be apparent to those of ordinary skill in the art in view of the disclosure herein are also within the scope of this invention. Accordingly, the scope of the invention is intended to be defined only by reference to the appended claims.

```
     <210> 25
     <211> 987
     <212> DNA
     <213> Homo sapiens
 5
     <220>
     <221> CDS
     <222> 238..609

10   <220>
     <221> sig_peptide
     <222> 238..291
     <223> Von Heijne matrix
           score 10.0374888212272
15         seq LLLLVMALPPGTT/GV


     <400> 25
     attccattca cagactcttg ttgggcagca gccacccgct cacctccatc cccaggactt      60
     agagggacgc agggcgttgg aacagagga cactccaggc gctgaccctg ggaggccagg     120
20   accagggcca aagtcccgtg ggcaagagga gtcctcagag gtccttcatt cagcggttcc     180
     gggaggtctg ggaagcccac ggcctggctg gggcagggtc aacgccgcca ggccgcc       237
     atg gtc ctg tgc tgg ctg ctg ctt ctg gtg atg gct ctg ccc cca ggc     285
     Met Val Leu Cys Trp Leu Leu Leu Leu Val Met Ala Leu Pro Pro Gly
               -15                       -10                      -5
25   acg acg ggc gtc aag gac tgc gtc ttc tgt gag ctc acc gac tcc atg     333
     Thr Thr Gly Val Lys Asp Cys Val Phe Cys Glu Leu Thr Asp Ser Met
              1                    5                    10
     cag tgt cct ggt acc tac atg cac tgt ggc gat gac gag gac tgc ttc     381
     Gln Cys Pro Gly Thr Tyr Met His Cys Gly Asp Asp Glu Asp Cys Phe
30   15                    20                   25                   30
     aca ggc cac ggg gtc gcc ccg ggc act ggt ccg gtc atc aac aaa ggc     429
     Thr Gly His Gly Val Ala Pro Gly Thr Gly Pro Val Ile Asn Lys Gly
                         35                   40                   45
     tgc ctg cga gcc acc agc tgc ggc ctt gag gaa ccc gtc agc tac agg     477
35   Cys Leu Arg Ala Thr Ser Cys Gly Leu Glu Glu Pro Val Ser Tyr Arg
                         50                   55                   60
     ggc gtc acc tac agc ctc acc acc aac tgc tgc acc ggc cgc ctg tgt     525
     Gly Val Thr Tyr Ser Leu Thr Thr Asn Cys Cys Thr Gly Arg Leu Cys
                         65                   70                   75
40   aac aga gcc ccg agc agc cag aca gtg ggg gcc acc acc agc ctg gca     573
     Asn Arg Ala Pro Ser Ser Gln Thr Val Gly Ala Thr Thr Ser Leu Ala
           80                   85                   90
     ctg ggg ctg ggt atg ctg ctt cct cca cgt ttg ctg tgaccaacag        619
     Leu Gly Leu Gly Met Leu Leu Pro Pro Arg Leu Leu
45   95                   100                  105
     ggaggacagg gcctgggact gttctcccag atccgccact ccccatgtcc ccatgtcctt     679
     cccccactaa atggccagag aggccctgga caacctcttg cggccctggc ttcatccctt     739
     ctaaggctgt ccaccaggag cccggtgcta ggggaagcat ccccaggcct gactgagcgg     799
     cagggggagca cggcccgtgg gtttgattgt attactctgt tccactggtt ctaagacgca     859
50   gagcttctca catctcaatc aggatgcttc tctccattgg tagcacttta gagtccatga     919
     aatatggtaa aaaatatata tatatcataa taaatgacag ctgatgttca tggaaaaaaa     979
     aaaaaaaa                                                              987


     <210> 26
55   <211> 908
     <212> DNA
     <213> Homo sapiens


     <220>
60   <221> CDS
     <222> 80..862

     <220>
```

```
      <220>
      <221> SIGNAL
      <222> -19..-1

  5   <400> 264
      Met Phe Leu Thr Val Lys Leu Leu Leu Gly Gln Arg Cys Ser Leu Lys
                      -15              -10                      -5
      Val Ser Gly Gln Glu Ser Val Ala Thr Leu Lys Arg Leu Val Ser Arg
                  1               5                   10
 10   Arg Leu Lys Val Pro Glu Glu Gln Gln His Leu Leu Phe Arg Gly Gln
          15              20                  25
      Leu Leu Glu Asp Asp Lys His Leu Ser Asp Tyr Cys Ile Gly Pro Asn
      30              35                  40                      45
      Ala Ser Ile Asn Val Ile Met Gln Pro Leu Glu Lys Met Ala Leu Lys
 15                   50                  55                      60
      Glu Ala His Gln Pro Gln Thr Gln Pro Leu Trp His Gln Leu Gly Leu
                  65                  70                  75
      Val Leu Ala Lys His Phe Glu Pro Gln Asp Ala Lys Ala Val Leu Gln
                  80                  85                  90
 20   Leu Leu Arg Gln Glu His Glu Glu Arg Leu Gln Lys Ile Ser Leu Glu
                  95                  100                 105
      His Leu Glu Gln Leu Ala Gln Tyr Leu Leu Ala Glu Glu Pro His Val
      110                 115                 120                 125
      Glu Pro Ala Gly Glu Arg Glu Leu Glu Ala Lys Ala Arg Pro Gln Ser
 25                   130                 135                 140
      Ser Cys Asp Met Glu Glu Lys Glu Glu Ala Ala Ala Asp Gln
                  145                 150                 155


      <210> 265
 30   <211> 106
      <212> PRT
      <213> Homo sapiens


      <220>
 35   <221> SIGNAL
      <222> -17..-1


      <400> 265
      Met Ala Leu Glu Val Leu Met Leu Leu Ala Val Leu Ile Trp Thr Gly
 40               -15                 -10                 -5
      Ala Glu Asn Leu His Val Lys Ile Ser Cys Ser Leu Asp Trp Leu Met
                  1               5                   10              15
      Val Ser Val Ile Pro Val Ala Glu Ser Arg Asn Leu Tyr Ile Phe Ala
                      20                  25                  30
 45   Asp Glu Leu His Leu Gly Met Gly Cys Pro Ala Asn Arg Ile His Thr
                      35                  40                  45
      Tyr Val Tyr Glu Phe Ile Tyr Leu Val Arg Asp Cys Gly Ile Arg Thr
                  .50                 55                  60
      Arg Val Arg Thr Val Ile Val Cys Lys Lys Tyr Cys Met Phe Cys Gln
 50       65                  70                  75
      Thr Phe Met Pro Ser Ile Lys Ile Val Phe
      80                  85


      <210> 266
 55   <211> 124
      <212> PRT
      <213> Homo sapiens


      <220>
 60   <221> SIGNAL
      <222> -18..-1

      <400> 266
```

251

```
       Met Val Leu Cys Trp Leu Leu Leu Leu Val Met Ala Leu Pro Pro Gly
                   -15                 -10               .           -5
       Thr Thr Gly Val Lys Asp Cys Val Phe Cys Glu Leu Thr Asp Ser Met
                    1           5                       10
   5   Gln Cys Pro Gly Thr Tyr Met His Cys Gly Asp Asp Glu Asp Cys Phe
       15              20                      25                      30
       Thr Gly His Gly Val Ala Pro Gly Thr Gly Pro Val Ile Asn Lys Gly
                       35                  40                  45
       Cys Leu Arg Ala Thr Ser Cys Gly Leu Glu Glu Pro Val Ser Tyr Arg
  10           50                  55                  60
       Gly Val Thr Tyr Ser Leu Thr Thr Asn Cys Cys Thr Gly Arg Leu Cys
                65                  70                  75
       Asn Arg Ala Pro Ser Ser Gln Thr Val Gly Ala Thr Thr Ser Leu Ala
           80                  85                  90
  15   Leu Gly Leu Gly Met Leu Leu Pro Pro Arg Leu Leu
       95                  100                 105
```

```
       <210> 267
       <211> 261
  20   <212> PRT
       <213> Homo sapiens

       <220>
       <221> SIGNAL
  25   <222> -16..-1

       <400> 267
       Met Glu Asn Phe Ser Leu Leu Ser Ile Ser Gly Pro Pro Ile Ser Ser
           -15                 -10                 -5
  30   Ser Ala Leu Ser Ala Phe Pro Asp Ile Met Phe Ser Arg Ala Thr Ser
        1           5                   10                  15
       Leu Pro Asp Ile Ala Lys Thr Ala Val Pro Thr Glu Ala Ser Ser Pro
                   20                  25                  30    .
       Ala Gln Ala Leu Pro Pro Gln Tyr Gln Ser Ile Ile Val Arg Gln Gly
  35           35                  40                  45
       Ile Gln Asn Thr Val Leu Ser Pro Asp Cys Ser Leu Gly Asp Thr Gln
               50                  55                  60
       His Gly Glu Lys Leu Arg Arg Asn Cys Thr Ile Tyr Arg Pro Trp Phe
       65                  70                  75                  80
  40   Ser Pro Tyr Ser Tyr Phe Val Cys Ala Asp Lys Glu Ser Gln Leu Glu
                       85                  90                  95
       Ala Tyr Asp Phe Pro Glu Val Gln Gln Asp Glu Gly Lys Trp Asp Asn
                   100                 105                 110
       Cys Leu Ser Glu Asp Met Ala Glu Asn Ile Cys Ser Ser Ser Ser Ser
  45           115                 120                 125
       Pro Glu Asn Thr Cys Pro Arg Glu Ala Thr Lys Lys Ser Arg His Gly
           130                 135                 140
       Leu Asp Ser Ile Thr Ser Gln Asp Ile Leu Met Ala Ser Arg Trp His
       145                 150                 155                 160
  50   Pro Ala Gln Gln Asn Gly Tyr Lys Cys Val Ala Cys Cys Arg Met Tyr
                       165                 170                 175
       Pro Thr Leu Asp Phe Leu Lys Ser His Ile Lys Arg Gly Phe Arg Glu
                   180                 185                 190
       Gly Phe Ser Cys Lys Val Tyr Tyr Arg Lys Leu Lys Ala Leu Trp Ser
  55           195                 200                 205
       Lys Glu Gln Lys Ala Arg Leu Gly Asp Arg Leu Ser Ser Gly Ser Cys
           210                 215                 220
       Gln Ala Phe Asn Ser Pro Ala Glu His Leu Arg Gln Ile Gly Gly Glu
       225                 230                 235                 240
  60   Ala Tyr Leu Cys Leu
                   245
```

```
       <210> 268
```

WHAT IS CLAIMED IS:

1.   An isolated polynucleotide, said polynucleotide comprising a nucleic acid sequence encoding:

        i)   a polypeptide comprising an amino acid sequence having at least about 80% identity to any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool; or

        ii)   a biologically active fragment of said polypeptide.

2.   The polynucleotide of claim 1, wherein said polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of the polypeptides encoded by the clone inserts of the deposited clone pool.

3.   The polynucleotide of claim 1, wherein said polypeptide comprises a signal peptide.

4.   The polynucleotide of claim 1, wherein said polypeptide is a mature protein.

5.   The polynucleotide of claim 1, wherein said nucleic acid sequence has at least about 80% identity over at least about 100 contiguous nucleotides to any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

6.   The polynucleotide of claim 1, wherein said polynucleotide hybridizes under stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

7.   The polynucleotide of claim 5, wherein said nucleic acid sequence comprises any one of the sequences shown as SEQ ID NOs:1-241 or any one the sequences of the clone inserts of the deposited clone pool.

8.   The polynucleotide of claim 1, wherein said polynucleotide is operably linked to a promoter.

9.   An expression vector comprising the polynucleotide of claim 8.

10.   A host cell recombinant for the polynucleotide of claim 1.

11.   A non-human transgenic animal comprising the host cell of claim 10.

12.   A method of making a GENSET polypeptide, said method comprising

      a) providing a population of host cells comprising the polynucleotide of
5   claim 8; and

      b) culturing said population of host cells under conditions conducive to the
             production of said polypeptide within said host cells.

13.   The method of claim 12, further comprising purifying said polypeptide from said
10 population of host cells.

14.   A method of making a GENSET polypeptide, said method comprising

      a) providing a population of cells comprising the polynucleotide of claim
             8;

15   b) culturing said population of cells under conditions conducive to the
             production of said polypeptide within said cells; and

    c) purifying said polypeptide from said population of cells.

15.   An isolated polynucleotide, said polynucleotide comprising a nucleic acid sequence
20 having at least about 80% identity over at least about 100 contiguous nucleotides to any one of the
sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the
deposited clone pool.

16.   The polynucleotide of claim 15, wherein said polynucleotide hybridizes under
25 stringent conditions to a polynucleotide comprising any one of the sequences shown as SEQ ID
NOs:1-241 or any one of the sequences of the clone inserts of the deposited clone pool.

17.   The polynucleotide of claim 15, wherein said polynucleotide comprises any one of
the sequences shown as SEQ ID NOs:1-241 or any one of the sequences of the clone inserts of the
30 deposited clone pool.

18.   A biologically active polypeptide encoded by the polynucleotide of claim 15.

19.   An isolated polypeptide or biologically active fragment thereof, said polypeptide
35 comprising an amino acid sequence having at least about 80% sequence identity to any one of the
sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by
the clone inserts of the deposited clone pool.

547

20.   The polypeptide of claim 19, wherein said polypeptide is selectively recognized by an antibody raised against an antigenic polypeptide, or an antigenic fragment thereof, said antigenic polypeptide comprising any one of the sequences shown as SEQ ID NOs:242-482 or any one of the

5   sequences of polypeptides encoded by the clone inserts of the deposited clone pool.

21.   The polypeptide of claim 19, wherein said polypeptide comprises any one of the sequences shown as SEQ ID NOs:242-482 or any one of the sequences of polypeptides encoded by the clone inserts of the deposited clone pool.

10

22.   The polypeptide of claim 19, wherein said polypeptide comprises a signal peptide.

23.   The polypeptide of claim 19, wherein said polypeptide is a mature protein.

15   24.   An antibody that specifically binds to the polypeptide of claim 19.

25.   A method of determining whether a GENSET gene is expressed within a mammal, said method comprising the steps of:

        a)   providing a biological sample from said mammal

20   b)   contacting said biological sample with either of:

                  i)   a polynucleotide that hybridizes under stringent conditions to the polynucleotide of claim 1; or

                  ii) a polypeptide that specifically binds to the polypeptide of claim 19; and

        c)   detecting the presence or absence of hybridization between said polynucleotide

25   and an RNA species within said sample, or the presence or absence of binding of said polypeptide to a protein within said sample;

wherein a detection of said hybridization or of said binding indicates that said GENSET gene is expressed within said mammal.

30   26.   The method of claim 25, wherein said polynucleotide is a primer, and wherein said hybridization is detected by detecting the presence of an amplification product comprising the sequence of said primer.

27.   The method of claim 25, wherein said polypeptide is an antibody.

35

28.   A method of determining whether a mammal has an elevated or reduced level of GENSET gene expression, said method comprising the steps of :

a)   providing a biological sample from said mammal; and

b)   comparing the amount of the polypeptide of claim 19, or of an RNA species
encoding said polypeptide, within said biological sample with a level
detected in or expected from a control sample;

5   wherein an increased amount of said polypeptide or said RNA species within said biological
sample compared to said level detected in or expected from said control sample indicates that said
mammal has an elevated level of said GENSET gene expression, and wherein a decreased amount
of said polypeptide or said RNA species within said biological sample compared to said level
detected in or expected from said control sample indicates that said mammal has a reduced level of

10   said GENSET gene expression.

29.   A method of identifying a candidate modulator of a GENSET polypeptide, said
method comprising :

a) contacting the polypeptide of claim 18 with a test compound; and

15   b) determining whether said compound specifically binds to said
polypeptide;

wherein a detection that said compound specifically binds to said polypeptide indicates that
said compound is a candidate modulator of said GENSET polypeptide.

20

549